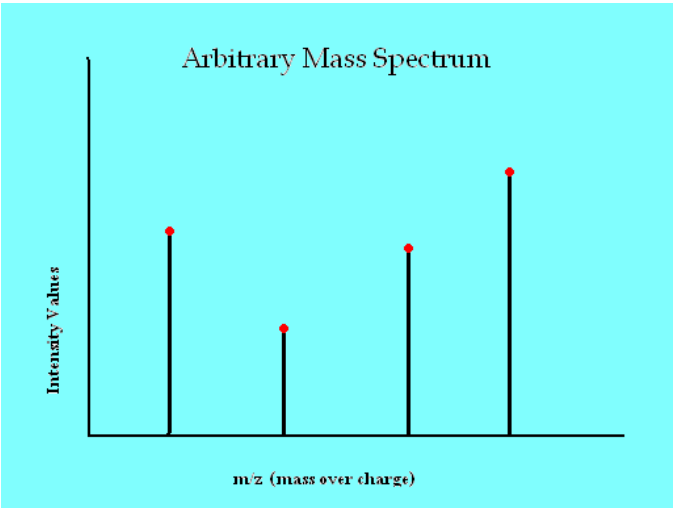
Inorganic and Organic Chemistry

ORGANIZING DATA IN MASS SPECTROMETRY

Barker, E. Brandon (Computer Science)
Jaromczyk, W. Jerzy* (Computer Science)
Staben, Chuck (Biology)

As the demand for identifying unknown compounds or measuring the amount of a certain compound in a chemical mixture increases, UK's Mass Spectrometry Facility has had to develop utilities enabling faster turnaround. Currently the facility offers services to characterize a wide variety of substances, including compounds originating from organic and organometallic synthesis, high molecular weight biological compounds such as peptides, proteins, oligonucleotides, complex lipids, and carbohydrates, as well as synthetic polymers. Although progress in hardware design enabling the collection of data allowed for improved processing speed, logistical improvements to every step of the analysis, including the collection, storage, and searching of data is still a critical factor. Difficulties occur when proprietary data formats, such as the data used by Mascot and Sequest databases, is needed in a single search. Since different spectrometers serve different roles, the variety of software associated with them must be used. This is not an ideal situation as often there is no convenient method to uniformly check one query in all of the mass spectra available.

While doing undergraduate research during the summer, I created a utility that converts a Mascot database file to a Sequest database entry. Each vendor's software stores data in a proprietary format, although the meaning of the data can be thought of as a graph like the figure to the left. My database converter from the Mascot database to the Sequest database can be viewed as a function where only the essential data is preserved, since the Mascot database stores many items not used by Sequest. Mascot and Sequest databases store data in flat text file and multi-flat text file formats, respectively. I selected Perl as the development language because of its superior ability to process large amounts of textual data. The results of this project enabled the spectrometry facility to substantially increase its throughput and response time for researchers in biology, chemistry, and medical fields. Previously a query would need to be submitted and processed on both databases, and substantial and cumbersome human-computer interaction was involved. Details of the implementation and further examples of mass spectra will be shown during the presentation.



Acknowledgements: Substantial help in understanding mass spectrometry was received from Dr. Jack Goodman and Dr. Bert Lynn.